Original Article

Trends in Hearing
Volume 25: 1–18
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/23312165211009351
journals.sagepub.com/home/tia
SAGE

Effect of Speech-to-Noise Ratio and Luminance on a Range of Current and Potential Pupil Response Measures to Assess Listening Effort

Patrycja Książek[1,2], Adriana A. Zekveld[1], Dorothea Wendt[2,3], Lorenz Fiedler[2], Thomas Lunner[2], and Sophia E. Kramer[1]

Abstract

In hearing research, pupillometry is an established method of studying listening effort. The focus of this study was to evaluate several pupil measures extracted from the Task-Evoked Pupil Responses (TEPRs) in speech-in-noise test. A range of analysis approaches was applied to extract these pupil measures, namely (a) pupil peak dilation (PPD); (b) mean pupil dilation (MPD); (c) index of pupillary activity; (d) growth curve analysis (GCA); and (e) principal component analysis (PCA). The effect of signal-to-noise ratio (SNR; Data Set A: –20 dB, –10 dB, +5 dB SNR) and luminance (Data Set B: 0.1 cd/m^2, 360 cd/m^2) on the TEPRs were investigated. Data Sets A and B were recorded during a speech-in-noise test and included TEPRs from 33 and 27 normal-hearing native Dutch speakers, respectively. The main results were as follows: (a) A significant effect of SNR was revealed for all pupil measures extracted in the time domain (PPD, MPD, GCA, PCA); (b) Two time series analysis approaches (GCA, PCA) provided modeled temporal profiles of TEPRs (GCA); and time windows spanning subtasks performed in a speech-in-noise test (PCA); and (c) All pupil measures revealed a significant effect of luminance. In conclusion, multiple pupil measures showed similar effects of SNR, suggesting that effort may be reflected in multiple aspects of TEPR. Moreover, a direct analysis of the pupil time course seems to provide a more holistic view of TEPRs, yet further research is needed to understand and interpret its measures. Further research is also required to find pupil measures less sensitive to changes in luminance.

Keywords

listening effort, pupillometry, principal component analysis, index of pupillary activity, cluster-based permutation test

Received 28 February 2020; Revised 11 February 2021; Accepted 5 March 2021

Hearing impairment has been shown to have a large effect on quality of life and cognitive decline in later life (Livingston et al., 2017). Hearing-impaired listeners tend to have difficulties with speech perception not only in terms of listening to a degraded speech signal but also in terms of processing and paying sustained attention to the speech (Holman et al., 2019; Koelewijn et al., 2015; Ohlenforst et al., 2017, 2018; Pichora-Fuller, 2003; Wingfield et al., 2015). Those difficulties affect daily functioning of people with hearing impairment and have been shown to be associated with increased need for recovery (Wang, Naylor, et al., 2018), high levels of effort-related fatigue (Holman et al., 2019), and mental distress leading to sick leave (Kramer et al., 2006). These adverse consequences of hearing impairment related to mental health are not fully addressed by using hearing aids (Holman et al., 2019; Keidser et al.,

[1]Amsterdam UMC, Vrije Universiteit Amsterdam, Otolaryngology-Head and Neck Surgery, Ear and Hearing, Amsterdam Public Health Research Institute, Amsterdam, the Netherlands
[2]Eriksholm Research Centre, Snekkersten, Denmark
[3]Department of Health Technology, Technical University of Denmark, Lyngby, Denmark

Corresponding author:
Patrycja Książek, Amsterdam UMC, Vrije Universiteit Amsterdam, Otolaryngology-Head and Neck Surgery, Ear and Hearing, Amsterdam Public Health Research Institute, Amsterdam, the Netherlands.
Email: p.ksiazek@amsterdamumc.nl

2015; Nachtegaal et al., 2009). A better understanding of how people perceive and process listening situations and which factors influence effort in adverse listening situations may be used to improve hearing care provided to hearing-impaired listeners. For example, the development of objective measures of listening effort could support hearing rehabilitation by providing tools to optimize speech perception performance while keeping listening effort low.

Pupillometry as an Objective Measure of Listening Effort

Listening effort is defined as "the deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a [listening] task" (Pichora-Fuller et al., 2016, p. 11S). It can be studied qualitatively using subjective ratings or interviews, or quantitatively with physiological measures, such as pupillometry. Pupillometry is a measure of temporal changes in pupil size. An increasing body of research has been focusing on the assessment of the transient increase in pupil size evoked by a task, called Task-Evoked Pupil Response (TEPR; review Zekveld et al., 2018). For example, assessment of the TEPR during speech perception has been applied often to examine listening effort (Koelewijn, Zekveld, Festen, & Kramer, 2012; Kramer et al., 1997; Kuchinsky et al., 2013; Ohlenforst et al., 2017; Wang, Naylor, et al., 2018; Wendt et al., 2018; Winn et al., 2015; Zekveld et al., 2010). The transient increase in pupil diameter in response to the stimulus (target speech) can reflect increasing listening effort until listening becomes too difficult. With further increasing difficulty levels, the TEPR decreases which has been interpreted as signs of disengagement (Ohlenforst et al., 2017, 2018; Wendt et al., 2018; Zekveld & Kramer, 2014).

Pupillometry is a noninvasive physiological measure of involuntary activity of the autonomous nervous system (ANS; Liu et al., 2017; Loewenfeld, 1999; Steinhauer & Hakerem, 1992; Steinhauer et al., 2004). However, as TEPRs seem to reflect multiple cognitive resources, arousal level, and emotional state, the interpretation of pupillometry results may be difficult (Francis & Oliver, 2018; Pichora-Fuller et al., 2016; review Zekveld et al., 2018). Furthermore, TEPRs are susceptible to external confounders such as luminance (Peysakhovich et al., 2017; Wang, Kramer, et al., 2018) or eye conditions such as anisocoria (Loewenfeld, 1999). In the current study, the effect of luminance on the eye was of special interest as this effect currently makes pupillometry applicable to lab settings only. First, because luminance effect may overshadow the effect of task demand on the TEPRs, as one of the main pupil functions is to regulate the amount of light entering the eye (Davson, 1990, p. 754). Second, because differences in the stimuli luminance level may influence the inner workings of the ANS, which does not always refer to the task demands (Wang et al., 2016, 2018). In addition, pupillometry results can be influenced by the steps taken to remove acquisition errors, cleaning and preprocessing data. Finally, the results depend on the specific measures that are extracted from the signal (Reilly et al., 2018; Winn et al., 2018). The latter factor was of interest in the current study.

In the current study, we were particularly interested in pupil measures that have been applied in several newer studies that aimed to test cognitive effort, preferably using acoustic stimuli presentation, and could be adopted in the audiological setting. These measures were either a single-value measure extracted to summarize the TEPR in the time or frequency domain, or multiple measures extracted from the time course of the TEPR (Kahneman & Beatty, 1966; Koelewijn, Zekveld, Festen, Rönnberg, et al., 2012; Wendt et al., 2018; Zekveld et al., 2010).

First, widely used single-value measures in the time domain include the pupil peak dilation (PPD) and mean pupil dilation (MPD) that aim to quantify the stimulus- or task-evoked dilation (Zekveld et al., 2018). The magnitude of dilation is calculated as a max or mean pupil size relative to a baseline obtained prior to the task. Second, single-value measures in the frequency domain include low frequency/high frequency ratio (LF/HF), index of cognitive activity (ICA), and index of pupillary activity (IPA). These aim to measure moment-to-moment pupil activity in response to cognitive effort (Boehm-Davis & Marshall, 2003; Duchowski et al., 2018; Peysakhovich et al., 2015). Besides, several ICA studies (Boehm-Davis & Marshall, 2003; Marshall, 2002) indicated that it is possible to separate dilation reflexes from light reflexes in the slightly preprocessed pupil data. This reported effectiveness in testing cognitive effort and filter out the effect of luminance at the same time seemed to be promising for testing listening effort in realistic scenarios. Third, the analysis of the pupil time course in form of a statistical test such as cluster-based permutation test or statistical model such as growth curve analysis (GCA), generalized additive mixed models (GAMMs), functional data analysis (FDA), or principal component analysis (PCA). The analyses aim to detect multiple time-dependent changes in TEPRs (Jackson & Sirois, 2009; Johansson et al., 2018; Mirman et al., 2008; van Rij et al., 2019). With statistical models, it is possible to quantify the pupil morphology (GCA, FDA, GAMM) or divide TEPRs into time windows (PCA). In the current study, examples of both single-value measures in time domain, frequency domain, and time course analysis were considered to provide a broad insight into pupil measures.

Single-Value Measures in the Time Domain—PPD and MPD. PPD and MPD are basic measures of TEPR and have been investigated in multiple studies within and outside of hearing research (Zekveld et al., 2018). Both PPD and MPD were used in studies investigating listening effort in a speech-in-noise or comprehension task (e.g., Ayasse et al., 2017; Ohlenforst et al., 2017, 2018; Wendt et al., 2018; Zekveld et al., 2010). An effect of signal-to-noise ratio (SNR) and speech intelligibility has been shown on the PPD measure (Ohlenforst et al., 2017; Zekveld et al., 2010), whereas an effect of age has been shown on the MPD measure (Ayasse et al., 2017). Greater PPD and MPD was interpreted as greater effort (Kahneman & Beatty, 1966; Zekveld et al., 2010). Audiological relevance and application of findings in studies using both PPD and MPD measures to investigate cognitive effort due to acoustic stimulation is well documented (review Zekveld et al., 2018). Thus, both measures were included in the current study.

Single-Value Measure in the Frequency Domain—Index of Cognitive/Pupillary Activity. ICA/IPA is a measure of abrupt changes in the pupil dilation accumulated within a time period, called *dilation reflexes*. Boehm-Davis and Marshall (2003) reported a higher ICA during difficult mental calculations compared with easy ones. They also reported no effect of luminance on the ICA measure. Demberg and Sayeed (2016) tested processing difficulty in a multitask paradigm that included listening as a secondary task to a driving task. ICA was significantly higher after a grammatical manipulation occurred in the auditory signal. Recently, Duchowski et al. (2018) partially replicated the findings of Boehm-Davis and Marshall (2003) by showing the effect of mathematical task difficulty on IPA in constant luminance. Each of the three studies described earlier reported an effect of task difficulty on ICA/IPA assumed to affect cognitive effort. They showed potential in quantifying cognitive effort not influenced by the luminance and findings were reproduced. If ICA/IPA reflect effort, we expected to find in this study the highest IPA/ICA values in TEPRs during speech-in-noise test with SNR levels corresponding to 50% speech intelligibility (Ohlenforst et al., 2017). On top of that, we expected no effect of luminance on ICA/IPA, meaning analyzing pupil responses in the frequency domain could be advantageous compared with analysis in the time domain where an effect of luminance was shown (Wang, Kramer, et al., 2018). In the current study, we included the IPA measure, which is inspired by ICA measures, and used the implementation reported by Duchowski et al.(2018). Unlike the original ICA implementation, IPA implementation has been made accessible to the researchers.

Analysis of the Pupil Time Course—Cluster-Based Permutation Test, GCA, and Principal Component Analysis. Time course analysis varies from less to more complex approaches. In the current study, three approaches with different complexity levels were compared: cluster-based permutation test, GCA, and PCA. The cluster-based permutation test aims to find clusters of samples significantly different between the conditions, where permutation limits the occurrences of false alarms (Johansson et al., 2018; Maris & Oostenveld, 2007; Sassenhagen & Draschkow, 2019). GCA is based on a polynomial function mimicking canonical TEPR (Kahneman & Beatty, 1966; Kuchinsky et al., 2013; Wendt et al., 2018), while PCA aims to divide the pupil response into time windows, without prior assumptions of the canonical TEPR, based on similarities between individual responses (Johansson et al., 2018; Zellin et al., 2011). As time course analysis is getting increasing attention in hearing research, both previously investigated (GCA) and relatively new to the field (cluster-based permutation test, PCA) approaches were included in the current study.

The cluster-based permutation test has been used to investigate listening effort in electroencephalography studies; however, examples of its use in pupillometry studies can be found as well (Dimitrijevic et al., 2019; Johansson et al., 2018; Maris & Oostenveld, 2007; Sassenhagen & Draschkow, 2019). Using the permutation test, Johansson et al. (2018) found that the effect of semantic coherence manipulation was present in the later stages of TEPR. Therefore, in the current study, we expected the effect of SNR and luminance to be reflected in a longer window of TEPRs, yet the length of clusters would differ between data sets.

GCA has been successfully implemented in multiple pupillometry studies (e.g., Kuchinsky et al., 2013; Wagner et al., 2019; Wendt et al., 2018; Winn, 2016). First, Kuchinsky et al. (2013) applied GCA to investigate differences in TEPR morphology due to lexical competition and background noise in a orthographic visual world paradigm test. The most challenging condition evoked the highest average response (intercept term), remained more elevated (linear term), more sustained (quadratic term), and had more delayed peak (cubic term) when compared with the less challenging conditions. Second, Wendt et al. (2018) applied GCA to investigate differences in morphology of the TEPRs in a speech-in-noise test, where the overall dilation was similar. They showed that TEPRs differed in height (intercept term) and sustain (quadratic term) between extreme SNR conditions, whereas the delay (cubic term) differed due to SNR in high intelligibility levels. On this basis, we expected in this study to see differences in the effects of SNR and luminance on GCA measures.

Several pupillometry studies on auditory perception have used PCA (Johansson et al., 2018; Zellin et al., 2011). Zellin et al. (2011) used PCA to evaluate the temporal changes of prosody perception and judgment in a set of short question-answer dialogs. They found three distinguishable components, where one was sensitive to the manipulation of the question's prosody. Differences in the scores of an active component seen 1–3 s after the answer's onset were associated with judgment of prosody adequacy. Johansson et al. (2018) used PCA to investigate the temporal changes in a process of memory retrieval manipulated by semantic coherence of the presented words to be remembered. The PCA identified three components. Two of them were sensitive to the coherence manipulation. Based on these previous studies, we expected that the TEPRs during speech-in-noise test would contain multiple components as a single trial usually is a few seconds long. Furthermore, we expected that at least one of the components would be affected by SNR and luminance similar to the other pupil measures.

Aims and Hypotheses

The current study aimed to answer three main questions: 1) Is there an effect of SNR on PPD, MPD, IPA, and measures obtained from GCA as well as from PCA? 2) What complementary information (if any) in describing the TEPR during speech-in-noise test do they provide? and 3) Is there an effect of luminance on any of the selected pupil measures? We compared a set of pupil measures potentially related to listening effort (PPD, MPD, IPA, GCA, and PCA) when applied to data acquired previously in a speech-in-noise test in two pupillometry studies (Ohlenforst et al., 2017; Wang, Naylor, et al., 2018). Based on the literature described earlier, we hypothesized that 1) all measures would show a significant effect of SNR, used as task difficulty manipulation; 2) GCA and PCA would provide multiple measures revealing the effect of SNR and/or luminance; and 3) in a data set with constant task demands (SNR) but varying luminance, no significant effect of luminance would be found on the IPA measures.

Methods

General Methods

The TEPRs were recorded in two studies in which participants had to listen to a target sentence and repeat it back. Those target sentences (Versfeld et al., 2000) were masked by a one-talker masker. Data Set A was collected by Ohlenforst et al. (2017), and Data Set B was collected by Wang et al. (2018). Ohlenforst et al. (2017) examined changes in PPD across a broad range of fixed SNRs, whereas Wang et al. (2018) assessed the effect of luminance on the PPD at 50% speech intelligibility. Both studies showed a significant effect of SNR and luminance, respectively, on the PPD measure. In both studies, TEPRs were recorded with a remote eye tracking system (SMI RED 500, SensoMotoric Instruments, Berlin, Germany) with a 60 Hz sampling rate.

Participants

Data Set A included pupil data from 33 normal-hearing (NH) native Dutch speakers (mean age = 47 years, *SD* = 12.1, ranging from 19 to 62). Data Set B included pupil data from 27 NH native Dutch speakers (mean age = 46 years, *SD* = 12.4, ranging from 21 to 58). The pure-tone hearing threshold average was obtained to check that the participants had NH. It was calculated from hearing threshold levels measured at each octave-band frequency between 500 and 4000 Hz in Data Set A and between 250 and 4000 Hz in Data Set B. Participants with a pure-tone hearing threshold average below 20 dB HL were included in the studies. Both studies were approved by the Ethics Committee of the VU University Medical Center in Amsterdam and performed at VU University Medical Center in Amsterdam. All participants provided written informed consent.

Task Characteristics

Data Set A was recorded during a speech-in-noise test using 9 fixed SNRs from –25 dB to +15 dB in 5-dB intervals, including 10 trials per SNR condition. The luminance level was adjusted individually so that the pupil size was an average between maximum in darkness and minimum in light (230 lux). Data Set B was recorded during a speech-in-noise test using two luminance conditions—darkness (0.1 cd/m^2) and light (360 cd/m^2). In both experiments, target sentences were presented to the participants binaurally via headphones (Sennheiser, HD 280). The masker level was constant and set to 65 dB sound pressure level. In Data Set A, the target sentence presentation level was fixed by design, and the order of SNR conditions was randomized per participant. In Data Set B, target presentation level varied in an adaptive procedure (one-up-one-down) to estimate the individual speech reception thresholds, that is, the 50% speech intelligibility. The first sentence in the list was presented at –10 dB SNR. This level was increased in 4 dB steps until the first sentence was repeated correctly. Further on, target sentence presentation level was increased or decreased in 2-dB steps, depending on the performance in the preceding trial (Wang, Naylor, et al., 2018). The speech reception threshold was calculated as the average target sentence presentation level in trials 5 to 25. Both data sets were recorded in studies using a blocked design with counterbalanced SNR and luminance conditions, respectively.

In the current study, three out of nine SNR conditions from Data Set A were selected to represent three distinguishable speech intelligibility (SI) levels, namely

–20 dB, –10 dB, +5 dB, corresponding to low (~5%) SI, medium (~50%) SI, and high (~95%) SI level. This reduced the number of pairwise comparisons without reducing the range of SI levels. Trials 1–4 from both luminance conditions in Data Set B were omitted as these trials were used to start the adaptive procedure. Figure 1 provides an overview of the timing of the trials for both data sets.

Preprocessing of the Pupil Data and Analysis Approaches

To fully use the selected analysis approaches, slightly different preprocessing stages were performed as shown in Table 1. Preprocessing and analysis approaches were applied using own preprocessing scripts in RStudio (Rstudio Team, 2016).

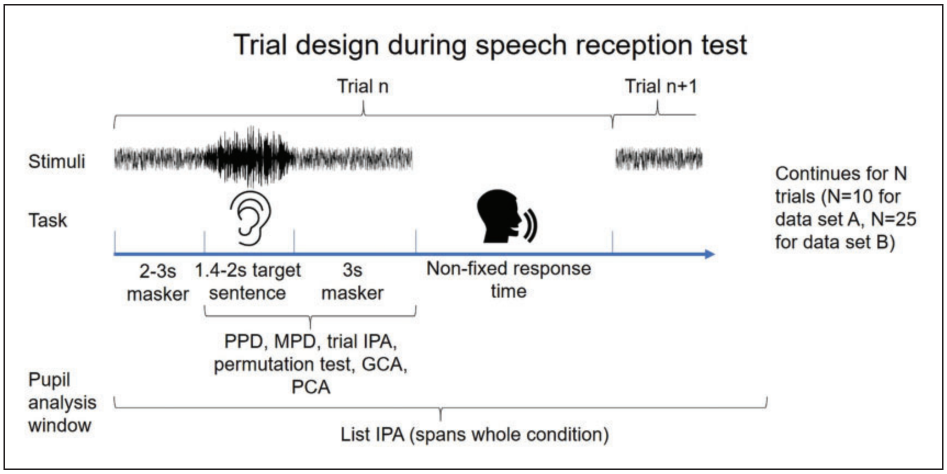
Trial design during speech reception test
Trial n
Trial n+1
Stimuli
Task
Continues for N trials (N=10 for data set A, N=25 for data set B)
2-3s masker
1.4-2s target sentence
3s masker
Non-fixed response time
PPD, MPD, trial IPA, permutation test, GCA, PCA
Pupil analysis window
List IPA (spans whole condition)

Figure 1. Schematic Trial Design During the Speech-in-Noise Test for Both Data Sets. Target sentence and masker were presented during a fixed time interval. In the response interval, participants were prompted to repeat the sentence. The response time was uncontrolled. The next trial was presented after scoring. Time was reset in the beginning of each trial. Artwork used in the scheme is licensed under CC BY-SA.
PPD = pupil peak dilation; MPD = mean pupil dilation; IPA = index of pupillary activity; GCA = growth curve analysis; PCA = principal component analysis.

Table 1. Preprocessing Steps Adjusted to the Used Analysis Approaches.

Analysis approach	Preprocessing of the signal
PPD, MPD, cluster-based permutation test, GCA and PCA	1. De-blinking of the raw data (83 ms before and 117 ms after each blink; Siegle et al., 2008), 2. Linear interpolation of the blinks, 3. Smoothing with a 7-point moving average filter, 4. Baseline correction at the trial level (mean 1-s long baseline prior to target sentence onset was subtracted from the TEPR), 5. Exclusion of the trials with >15% of missing data, 6. Exclusion of the data of the participants with >15% of excluded trials per condition. In Data Set A, the data of *n* = 0 participants were excluded in each of the tested SNRs. In Data Set B, the data of *n* = 2 participants were excluded in dark and light conditions, and 7. Averaging across the trials per participant and per condition.
IPA	1. De-blinking of the raw data as for PPD and PCA, 2. Linear interpolation of the blinks to increase continuity of the pupil response, 3. Exclusion of a. List IPA (Figure 1): Conditions with >30% of missing data in Data Set A and > 30% of missing data in Data Set B. In Data Set A, pupil data from *n* = 0 participants were excluded. In Data Set B, pupil data from *n* = 3 participants in dark and *n* = 0 participants in light condition. b. Trial IPA: The trials with >15% of missing data, and c. Exclusion of the data from participants with >15% of excluded trials per condition. In Data Set A, no participant was excluded in any of the conditions. In Data Set B, *n* = 2 participants were excluded in the dark and light conditions.

Note. PPD = pupil peak dilation; MPD = mean pupil dilation; GCA = growth curve analysis; PCA = principal component analysis; TEPR = Task-Evoked Pupil Response; SNR = signal-to-noise ratio; IPA = index of pupillary activity.

Pupil Peak Dilation

The PPD is the maximum value of the TEPR relative to the trial baseline (see Figures 1 and 2 for the baseline period) and was calculated as in the original studies (Ohlenforst et al., 2017; Wang, Naylor, et al., 2018). PPD values were extracted from preprocessed data and averaged for each condition and participant TEPRs (Table 1).

Mean Pupil Dilation

The MPD is the average TEPR relative to the trial baseline. Similarly to the other approaches, MPD was calculated from the 5-s long time window (see Figures 1 and 2). MPD values were extracted from preprocessed data and averaged for each condition and participant TEPR data (Table 1).

Index of Pupillary Activity

The IPA calculation was adapted from Duchowski et al. (2018) and performed in RStudio with a package wavelets (Aldrich, 2020; Percival & Walden, 2000). Wavelet decomposition was performed at a second level of resolution, which allowed to investigate the dilation reflexes (wavelet coefficients) between 7.5 and 15 Hz. We used wavelet function Daubechies-4, which allowed us to target dilations at ~7.8 Hz (i.e., 8 effective coefficients spanning a ~128 ms window as recorded with 60 Hz sampling frequency). A hard threshold with universal function Λ (Lambda) was applied to reduce the number of nonrelevant reflexes.

In this study, the IPA was calculated in two different ways. First, a list IPA was calculated for concatenated TEPRs within each condition (10 trials per SNR condition for Data Set A and 25 trials per luminance condition for Data Set B). Note that the list duration differed across participants due to variable response time, variable scoring interval durations (see Figure 1), a varying number of trials retained during data preprocessing (Table 1), and across data sets due to differences in masker noise duration. Second, a trial IPA was calculated within a fixed 5-s analysis window (see Figure 1), which spanned only the target sentence and masker presentation, and trial IPAs were then averaged across trials per condition and participant. Figure 3 illustrates an example of detected trial and list IPA.

Cluster-Based Permutation Test

The cluster-based permutation test was performed on preprocessed (see Table 1), averaged (per participant and per condition), 5-s long TEPRs (see Figure 1) in each data set separately. The test was implemented in RStudio with a package permuco (Frossard & Renaud, 2019). The cluster-based permutation test was used to test for the effects of SNR and luminance on TEPRs; however, it did not provide pupil measures comparable with the other approaches or across data sets

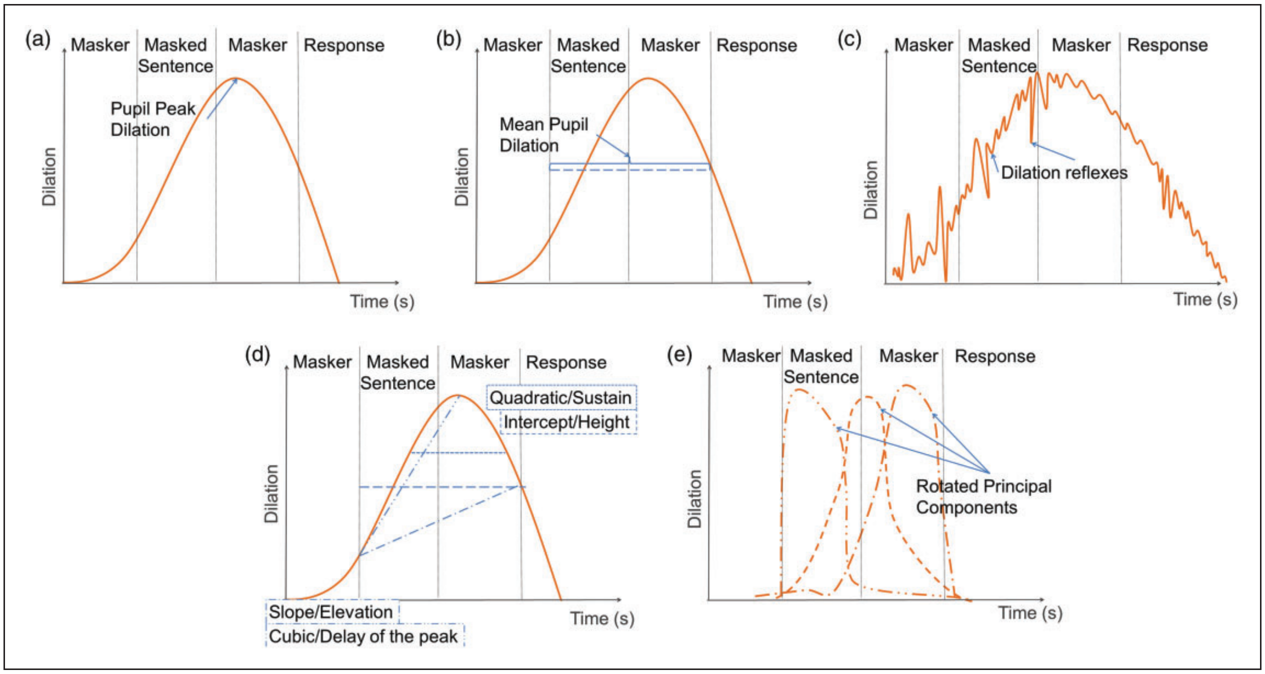

Figure 2. Schematic Overview of the Analysis Approaches. (a) Pupil peak dilation, (b) Mean pupil dilation, (c) Index of pupillary activity, (d) Growth curve analysis, and (e) principal component analysis, applied to TEPR recorded during speech-in-noise perception testing.

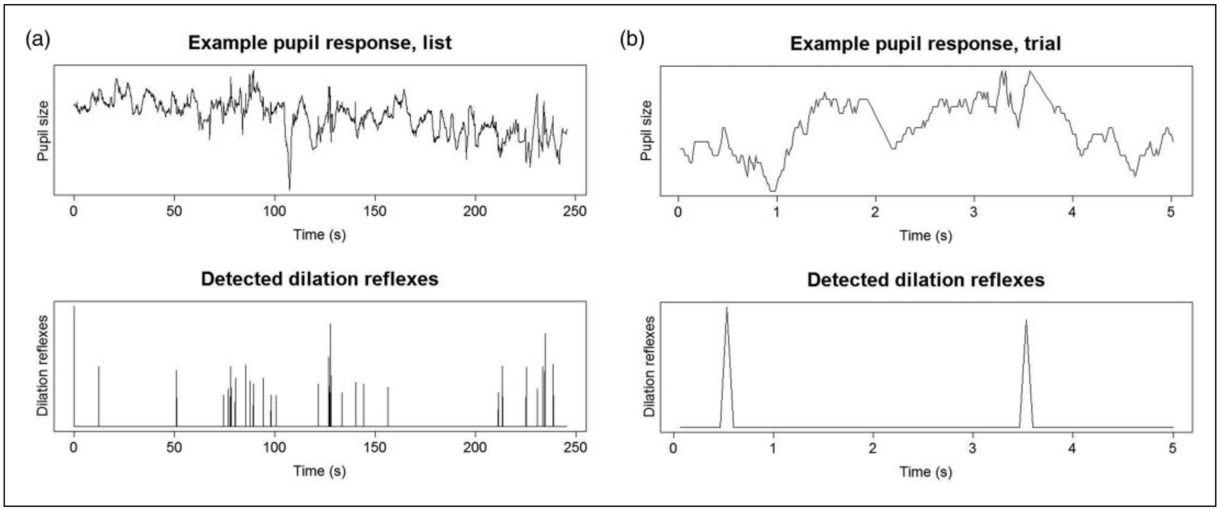

Figure 3. Example of dilation reflexes detected through wavelet decomposition. Detected dilation reflexes were used to calculate (a) list IPA and (b) trial IPA.

(Sassenhagen & Draschkow, 2019). Thus, this approach was not included in Figure 2.

Growth Curve Analysis

GCA with a polynomial as a functional basis was fitted to TEPRs for each data set separately. GCA model was fitted to five seconds long TEPRs averaged across trials per participant in each tested condition (see Figure 1). Following previous studies (Książek, 2017; Wendt et al., 2018), the GCA model was defined as an interaction between the polynomial terms (intercept, linear, quadratic, cubic) and SNR/luminance as a fixed factor. Higher order polynomial terms (quartic, quantic) were not included as we expected the main dilation 2–3 s after sentence onset (Wendt et al., 2018; Winn et al., 2018), were not interested in the secondary peak, and did not want to risk curve overfitting. The number of the polynomial terms related to the random factor was selected based on the χ^2 statistics. The fitting and selection process was done in RStudio with use of the lme4 package (Bates et al., 2015). As an outcome of the fitting process, third-order polynomial with four fixed and four random terms—intercept, linear, quadratic, and cubic—was selected in both data sets (see Figure 2 for graphical representation of the GCA measures). Diagnostics for GCA models are reported in Supplementary Material 1 (for selecting number of random factors only). Model fit is reported in Supplementary Material 2. The selected model was identical to the one used in Wendt et al. (2018).

Principal Component Analysis

A temporal PCA was performed on the averaged TEPRs from each tested condition (average across trials, separately for each per participant). For both data sets, the same 5-s time window starting at target sentence onset (Figure 1) was analyzed in three steps: (1a) identification of subgroups of pupil data points that were highly correlated with each other, which constructed the principal components (PCs) in an unrestricted PCA (Bishop, 2006; Johansson et al., 2018; Kayser & Tenke, 2003; Zellin et al., 2011); (1b) retaining PCs until the sum of explained variance in the data reached at least 80% (Stevens, 1996) or until the next retained PC would not add more to the amount of explained variance as indicated by a scree plot (Costello & Osborne, 2005); (2) transforming retained PCs in a rotated principal components (RPCs) analysis through varimax rotation with Kaiser normalization to increase their interpretability (Johansson et al., 2018; Kayser & Tenke, 2003; Verney et al., 2004; Wetzel et al., 2016); and (3) obtaining RPC scores as a measure of individual contribution by multiplying RPC loadings by the input pupil responses to enable comparisons between conditions. PCA was performed in R with use of two functions: function principal from package psych (Revelle, 2016) and a built-in function prcomp (and varimax for varimax rotation) from package stats (R Core Team, 2020). Standardized RPC loadings and approximate variance explained were plotted and approximated with function principal; RPC scores were obtained with function prcomp (+ varimax).

Evaluation of the Pupil Measures

Linear Mixed-Effects Models (LMMs) were selected to test and compare the effects of SNR (–20, –10, and +5 dB) and luminance (dark, light) on the pupil measures (PPD, MPD, GCA, RPC scores, IPA) that were extracted per participant and per SNR/luminance condition. LMMs were applied to account for differences in variance of pupil measures, unbalanced number of examples per condition, individual differences, and/or to correct for the autocorrelation. Several LMMs were fitted to each pupil measure and each data set separately in RStudio with the lme4, nlme, and stats packages (Bates et al., 2005; Pinheiro et al., 2020). Four models were fitted to all pupil measures (with exception of GCA) within a data set. Model 1—LM with SNR/luminance as a fixed factor and no random factors; Model 2—LMM with SNR/luminance as a fixed factor and varying intercept per participant as a random factor; Model 3—LMM with SNR/luminance as a fixed factor, varying intercept and slope per participant as random factors; Model 4—LMM with SNR/luminance as a fixed factor, varying intercept per participant as a random factor and autoregressive term affecting auto-correlated structure of time series (van Rij et al., 2019). Best-fitting model was selected based on four criteria: convergence, distribution of residuals, Akaike information criterion (Wagenmakers & Farrell, 2004), and estimation confidence intervals. Model diagnostics and selection are described in detail in Supplementary Material 1. An exception was GCA, where the pupil measures constituted the fixed effects that were estimated by a longitudinal LMM. Thus, in this case, fixed and random effects of SNR and luminance were estimated on the third polynomial as a functional basis instead of the intercept (described in the Growth Curve Analysis section).

To test for the main effect of SNR, an LMM analysis of variance (ANOVA) was performed on the best-fitting model per pupil measure and data set. Besides, to test for the main effect of luminance and differences between SNR conditions, Tukey's all pairwise comparisons were performed using the package multcomp (Hothorn et al., 2008). Tukey's all pairwise comparisons compensate for differences in the estimation procedures between models from the various packages that were applied. In case Model 1 showed the best fit, a *t* test was performed in the pairwise comparisons. For Models 2, 3, and 4 that included random factors, a *z* test was performed. The additive estimated effects of SNR and luminance on the GCA measures were evaluated through a *t* test with use of the lmerTest package (Kuznetsova et al., 2017). Multiple *t* tests were corrected for Type III error by Satterwaite's approximation degrees of freedom.

Results

First, pupil data were preprocessed and the TEPRs plotted to reproduce the previous results (Ohlenforst et al., 2017; Wang, Kramer, et al., 2018). The average responses are shown in Figure 4. For both Data Sets A and B, the TEPRs were comparable to the ones reported previously. All TEPRs were aligned with the masker onset (indicating beginning of the trial).

Cluster-Based Permutation Test

In Figure 4, we present the results of the cluster-based permutation test. These refer directly to clusters significantly different due to SNR and luminance. Note that the cluster-based permutation test did not provide specific pupil measures but indicated which clusters significantly differed between task conditions. The TEPR differed significantly between SNR conditions in the cluster between 0.32 and 5 s—F(2, 64) = 3.14, nr. Perm = 5000, cl. Mass = 3578.01, *P*(>mass) = .0002—in Data Set A.

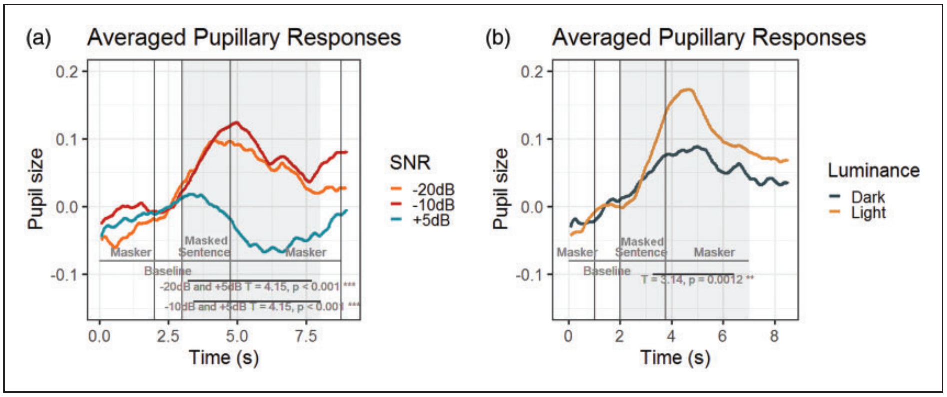

Figure 4. The grand averaged Task-Evoked Pupil Responses (TEPRs) recorded in the data sets. The averaged pupillary responses, that is Task-Evoked Pupil Responses (TEPRs), recorded in the two data sets. Averaged TEPRs for the (a) SNR and (b) luminance conditions. Shadowed region represents the analysis time window. This included masked target sentence presentation and following masker alone presentation. TEPRs were aligned with the masker onset. The data during the presentation of the masker in the 1-second interval before sentence onset were used for the baseline correction (2s-3s in (a), 1s-2s in (b)). Time-coded horizontal lines show the clusters in which the cluster-based permutation test revealed the effect of (a) SNR and (b) luminance.
SNR = signal-to-noise ratio.

Significant differences between luminance conditions were observed between 1.27 s and 4.43 s—F(1,30) = 4.17, nr. Perm = 5000, cl. Mass = 2085.29, *P*(>mass) = .0012—in Data Set B. Furthermore, pairwise comparisons indicated that the TEPR differed between the SNR conditions –20 dB and +5 dB SNR between 0.2 s and 4.7 s—F(1,32) = 4.15, nr. Perm = 5000, cl. Mass = 3783.19, *P* (>mass) = .0004. Also, the TEPR differed between the –10 dB and +5 dB SNR conditions between 0.42 s and 5 s—F(1,32) = 4.15, nr. Perm = 5000, cl. Mass = 5962.14, *P*(>mass) = .0002. The test revealed no significant differences between the –20 dB and –10 dB SNR conditions.

Pupil Peak Dilation

Figure 5 presents the PPD across conditions for Data Sets A and B. The highest PPD was found in –10 dB SNR and lowest in +5 dB SNR. LMM ANOVA revealed a significant main effect of SNR, *F*(2, 165) = 17.064, $p < .001$, and luminance, *F*(1, 87) = 35.229, $p < .001$, on PPD. The result of the pairwise comparisons as well as the model fits and β estimates are reported in Supplementary Material 2.

Mean Pupil Dilation

Figure 6 presents the MPD across conditions for Data Sets A and B. The median MPD was highest in the –10 dB SNR condition. LMM ANOVA revealed a significant main effect of SNR, *F*(2, 165) = 30.75, $p < .001$, and luminance, *F*(1, 87) = 19.947, $p < .001$, on MPD. Model fits, β estimates and pairwise comparisons are reported in Supplementary Material 2.

Index of Pupillary Activity

Figure 7 presents list and trial IPA extracted from Data Sets A and B. The overall IPA calculated per trial was

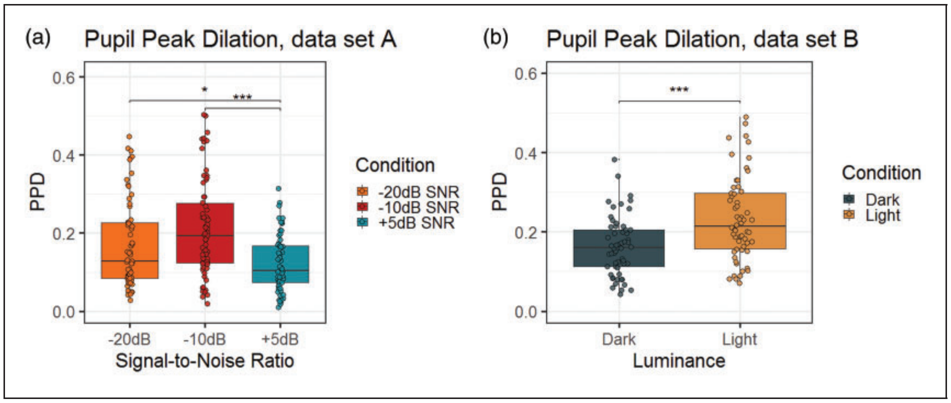

Figure 5. Visual representation of PPD statistics (boxplot) overlaid with the distribution of individual PPDs (jittered dots). PPDs are presented as a function of (a) SNR condition, Data Set A and (b) luminance condition, Data Set B. Statistical results of Tukey's all pairwise comparisons: **p*<.05, ****p*<.001.
SNR = signal-to-noise ratio; PPD = pupil peak dilation.

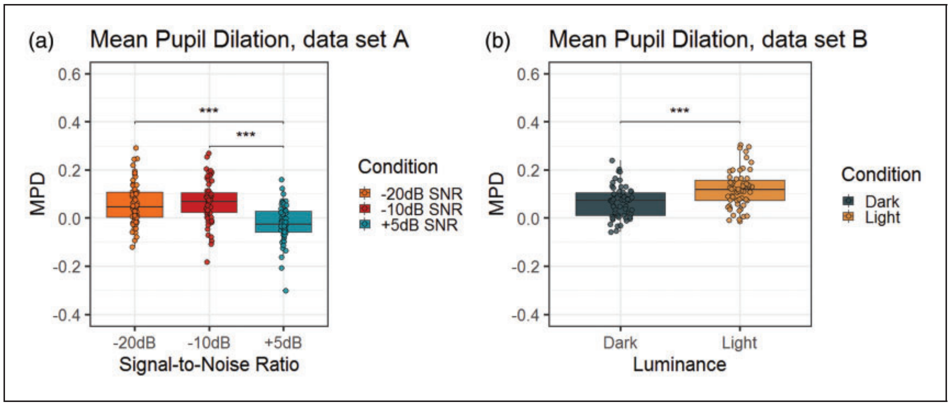

Figure 6. Visual representation of MPD statistics (boxplot) overlaid with the distribution of individual MPDs (jittered dots). MPDs are presented as a function of (a) SNR condition, Data Set A and (b) luminance condition, Data Set B. Statistical results of Tukey's all pairwise comparisons: ****p*<.001.
SNR = signal-to-noise ratio; MPD = mean pupil dilation.

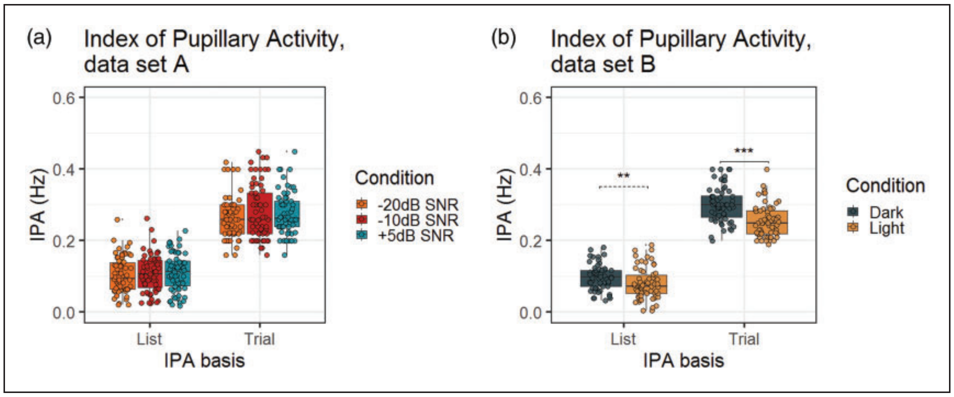

Figure 7. Visual representation of IPA statistics (boxplot) overlaid with the distribution of individual IPAs (jittered dots). IPA are presented as a function of (a) SNR condition, Data Set A and (b) luminance condition, Data Set B. Statistical results of Tukey's all pairwise comparisons: $^{**}p<.01$, $^{***}p<.001$.
SNR = signal-to-noise ratio; IPA = index of pupillary activity.

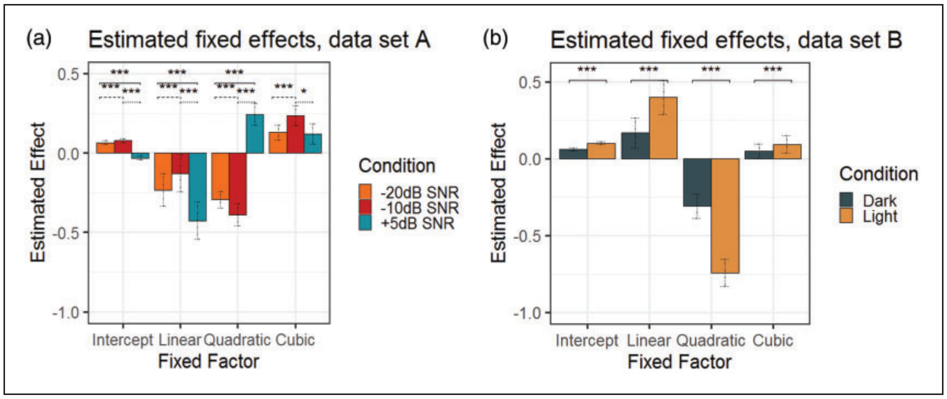

Figure 8. Visual representation of the GCA results (estimated polynomial terms) together with the standard errors (barplot). Data are presented for the (a) SNR condition, Data Set A and (b) luminance condition, Data Set B. Statistical results of Tukey's all pairwise comparisons: $^{*}p<.05$, $^{***}p<.001$.
SNR = signal-to-noise ratio.

higher than the one calculated per list, which was likely affected by the length of the input signal to the wavelet decomposition—denominator in the IPA formula.

LMM ANOVA revealed no effect of SNR on any of the IPA measures in Data Set A. However, the effect of luminance was found on both trial IPA, $F(1, 92.5) = 36.024$, $p < .001$, and on the list IPA, $F(1, 90.6) = 11.363$, $p = .0011$, Model fits, β estimates and pairwise comparisons are reported in the Supplementary Material 2.

Growth Curve Analysis

Figure 8 presents the estimated effect of SNR and luminance on GCA measures in Data Sets A and B, respectively. In Data Set A, the intercept (mean), quadratic term (elevation) as well as the cubic term (delay) were strongest for the –10 dB SNR condition as indicated by the estimated GCA measures. Note that the relatively large quadratic term in the –10 and –20 dB SNR conditions indicate a less sustained TEPR. The positive value of the quadratic term in the +5 dB SNR condition indicates that this term reflects the reduction in pupil size following the peak of the TEPR. Each of the GCA terms was stronger for the light when compared with the dark condition in Data Set B.

LMM ANOVA revealed a main effect of SNR on the intercept, $F(2, 59268) = 8427.028$, $p < .001$; linear, $F(2, 59268) = 176.759$, $p < .001$; quadratic, $F(2, 59268) = 901.286$, $p < .001$; and cubic, $F(2, 59268) = 31.19$, $p < .001$ term. The effect of luminance was revealed on the intercept, $F(1, 37076) = 4781.28$, $p < .001$; linear, $F(1, 37076) = 496.28$, $p < .001$; quadratic, $F(1, 37076) = 1746.06$, $p < .001$; and cubic, $F(1, 37076) = 17.42$, $p < .001$, term. Model fits and additive effects on the GCA measures are reported in Supplementary Material 2.

Table 2. Activity Time and Approximate Variance Explained Per RPC (Values Obtained With Function Principal).

Rotated principal components	Data Set A			Data Set B		
	RPC1	RPC2	RPC3	RPC1	RPC2	RPC3
Variance explained (%)	36.2	27.3	29.2	33.8	24.5	32.2
Activity time relative to sentence onset (ms)	3000–4850	3983–6683	5467–8017	2000–3750	3133–5667	4850–7017

Note. RPC = rotated principal component.

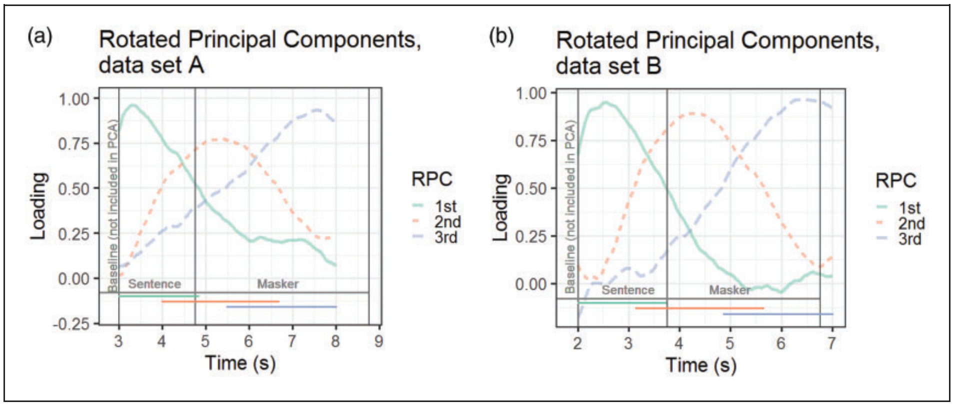

Figure 9. Activity time of the identified and retained by PCA Rotated Principal Components (loadings). Presented RPC loadings in (a) Data Set A and (b) Data Set B were standardized for easier comparison between the data sets. A 5-s long analysis window starting at target sentence onset was used for PCA. Lines at the bottom indicate activity time for each RPC.
RPC = rotated principal component; PCA = principal component analysis.

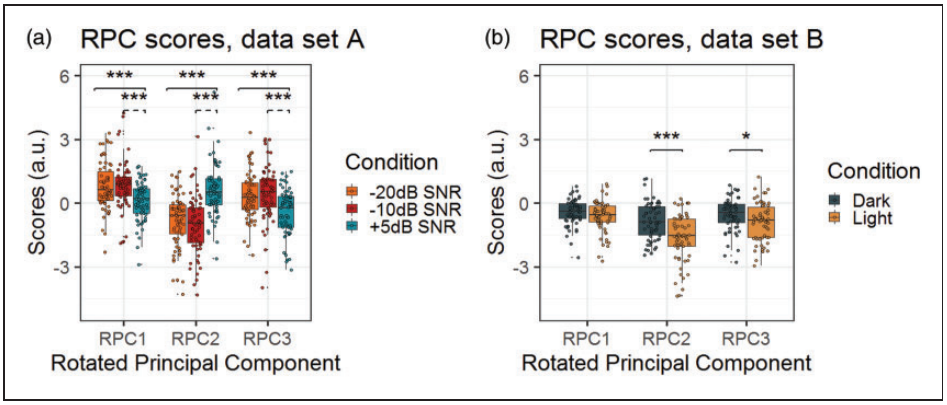

Figure 10. Visual representation of the RPC scores' statistics (boxplot) overlaid with the distribution of individual RPC scores (jittered dots). RPC scores are presented for (a) SNR condition, Data Set A and (b) luminance condition, Data Set B. Statistical results of Tukey's all pairwise comparisons: $*p<.05$, $***p<.001$.
RPC = rotated principal component; SNR = signal-to-noise ratio

Principal Component Analysis

In both data sets, three RPCs were obtained. In Data Set A, the three RPCs explained 92.7% of variance, whereas in Data Set B, they explained 90.5% of variance. Variance explained by each of the RPCs is presented in Table 2. Figure 9 shows the RPCs' main activity time (i.e., RPC loadings > .50) and their loadings as function of time. Each of the three RPCs spanned approximately the same time window in both data sets. In both data sets, the activity time overlapped across RPCs. Note that RPC1 was mostly active in the interval in which the masked target sentence was presented, and RPC3 was mostly active during the presentation of the masker after target sentence offset. RPC2 was mostly active in between RPC1 and RPC3, at the end of the target sentence and during the following masker presentation.

Figure 10 presents the RPC scores for Data Sets A and B. LMM ANOVA performed on β estimates revealed a main effect of SNR on the RPC1 scores, $F(2, 165) = 18.82$, $p < .001$; the RPC2 scores, $F(2, 163) = 39.85$, $p < .001$; and the RPC3 scores, $F(2, 165) = 18.71$, $p < .001$, and a main effect of luminance on the RPC2 scores, $F(1, 122) = 20.04$, $p < .001$, and the RPC3 scores, $F(1, 122) = 4.95$, $p = .03$. All pairwise comparisons on the RPC scores as well as the information about the model fits and β estimates are gathered in Supplementary Material 2. Taken together, the RPC loadings show that the listening task in the speech-in-noise test can be divided into three different stages independently of the tested condition. SNR affected the scores of each RPC, while the effect of luminance was found on RPC2 and RPC3 scores.

Summary of the Results

The implementation complexity differed between the approaches used to acquire the pupil measures assessed in this study. Furthermore, the measures differed in the nature and detail of the information provided about the effect of the conditions on the TEPRs. Besides, the LMM model that provided the best fit differed between several pupil measures. A summary of the best-fitting models is provided in Table 3. As shown, many of the single-value pupil measures (MPD, List IPA) could be explained with models including only the fixed factors SNR and luminance. In contrast, the best-fitting models of the time course analysis (GCA, PCA) measures also

Table 3. Selected Best-Fitting Model for Each Pupil Measure.

	RPC1	RPC2	RPC3	PPD	MPD	IPA trial	IPA list	GCA
Data Set A	Model 2	Model 4	Model 2	Model 3	Model 1	Model 2	Model 1	Model 5
Data Set B	Model 3	Model 1	Model 1	Model 2	Model 1	Model 2	Model 1	Model 5

Note. Model 1—LM: Pupil measure $\sim$ SNR/luminance; Model 2—LMM Pupil measure $\sim$ SNR/luminance + (1|Subject); Model 3—LMM Pupil measure $\sim$ SNR/luminance + (1+SNR/luminance|Subject); Model 4—LMM with SNR/luminance + (1|Subject) + corCAR; Model 5—LMM Pupil response $\sim$ (1 + Linear + Quadratic + Cubic) $\times$ SNR/luminance + (1 + Linear + Quadratic + Cubic |Subject). RPC = rotated principal component; PPD = pupil peak dilation; MPD = mean pupil dilation; IPA = index of pupillary activity; GCA = growth curve analysis.

Table 4. Comparison of Effects of SNR and Luminance Detected With the Investigated Pupil Measures.

	Data Set A			Data Set B
	–20 dB and –10 dB (<5% and 50% perform. Low-Middle)	–20 dB and +5 dB (<5% and >95% perform. Low-High)	–10 dB and +5 dB (50% and >95% perform. Middle-High)	Effect of luminance
Cluster-based permutation test		Cluster 3.2–7.7 s, $P(>\text{mass}) < .001$***	Cluster 3.42–8 s, $P(>\text{mass}) < .001$***	Cluster 3.27–6.43 s, $P(>\text{mass}) = .0012$**
PPD		$\Delta\beta = –0.048$, $p = .04$*	$\Delta\beta = –0.09$, $p < .001$***	$\Delta\beta = 0.069$, $p < .001$***
MPD		$\Delta\beta = –0.081$, $p < .001$***	$\Delta\beta = –0.09$, $p < .001$***	$\Delta\beta = 0.053$, $p < .001$***
Trial IPA				$\Delta\beta = –0.041$, $p = < .001$***
List IPA				$\Delta\beta = –0.019$, $p = .007$**
GCA Int	$\Delta\beta = 0.012$, $p < .001$***	$\Delta\beta = –0.097$, $p < .001$***	$\Delta\beta = –0.109$, $p < .001$***	$\Delta\beta = 0.041$, $p < .001$***
GCA Lin	$\Delta\beta = 0.103$, $p < .001$***	$\Delta\beta = –0.195$, $p < .001$***	$\Delta\beta = –0.283$, $p < .001$***	$\Delta\beta = 0.231$, $p < .001$***
GCA Quad	$\Delta\beta = –0.095$, $p < .001$***	$\Delta\beta = 0.538$, $p < .001$***	$\Delta\beta = 0.634$, $p < .001$***	$\Delta\beta = –0.433$, $p < .001$***
GCA Cub	$\Delta\beta = 0.106$, $p < .001$***		$\Delta\beta = –0.092$, $p = .02$*	$\Delta\beta = 0.043$, $p < .001$***
RPC1 (A: 3–4.85 s, B: 2–3.75 s)		$\Delta\beta = –0.719$, $p = < .001$***	$\Delta\beta = –0.767$, $p = < .001$***	
RPC2 (A: 3.98–6.68 s, B: 3.13–5.67 s)		$\Delta\beta = 1.534$, $p < .001$***	$\Delta\beta = 1.642$, $p = < .001$***	$\Delta\beta = –0.803$, $p < .001$***
RPC3 (A: 5.47–8 s, B: 4.85–7 s)		$\Delta\beta = –0.802$, $p < .001$***	$\Delta\beta = –0.934$, $p < .001$***	$\Delta\beta = –0.331$, $p = .03$*

Note. Statistical tests performed as followed: (a) Fisher *t* test with 5,000 permutations (cluster-based permutation test); (b) Tukey's all pairwise *t* test or *z* test (PPD, MPD, IPA, RPC scores); (c) Estimated β-coefficients for additive effects with multiple *t* test (GCA measures). Significant results are marked as follows: *$p < .05$. **$p < .01$. ***$p < .001$. PPD = pupil peak dilation; MPD = mean pupil dilation; IPA = index of pupillary activity; GCA = growth curve analysis; RPC = rotated principal component.

included a random intercept (PCA) and higher order polynomial terms (GCA, PCA).

As seen in Table 4, the significant effect of luminance was observed in all investigated pupil measures in Data Set B, which disagrees with Hypothesis 3, stating that IPA measures would not reveal this effect. The effect of SNR was observed on all pupil measures derived from the time domain (PPD, MPD, GCA, PCA); however, the additive effects of SNR varied across the pupil measures in Data Set A.

Discussion

The Effect of SNR and Luminance on the Pupil Measures

In the current study, a set of pupil measures (PPD, MPD, IPA, GCA, and PCA) potentially related to listening effort was evaluated in terms of their sensitivity to changes in SNR and luminance. The selected measures were derived from both the time and frequency domain of the TEPR. Moreover, the measures included single-value measures as well as measures that were estimated through a time course analysis. The TEPRs were acquired during a speech-in-noise test in two different pupillometry studies (Ohlenforst et al., 2017; Wang, Naylor, et al., 2018). More specifically, this study investigated the hypotheses that SNR as a manipulation of task demand would significantly affect all measures (Hypothesis 1) and that luminance would significantly affect all pupil measures except IPA (Hypothesis 3). In addition, it was hypothesized that multiple measures derived from the GCA and PCA analyses would be sensitive to the task manipulations (Hypothesis 2). Testing these hypotheses using several pupil measures not only enabled a thorough test of the effects of SNR and luminance but also highlighted the more specific benefits and drawbacks of the measures.

With respect to Hypothesis 1, the study revealed a significant effect of SNR, corresponding to 0%, 50%, and 95% speech intelligibility, on all pupil measures derived in the time domain (PPD, MPD, GCA, and PCA), but not on the measure derived in the frequency domain (IPA). This suggests that these measures may all tap into the effects of auditory task demand (listening effort). Only GCA revealed a significant effect of SNR at low intelligibility levels, that is, 0% versus 50% intelligibility and thereby provided more information about the effect of SNR on the TEPR than the single-value measures, which was in agreement with Hypothesis 2. In disagreement with Hypothesis 3, a significant effect of luminance was found for the pupil measures extracted in the frequency domain (IPA).

Descriptive Comparison of the Analysis Approaches

Single-Value Measures in the Time Domain. The effect of SNR and luminance was revealed on both PPD and MPD. This replicates findings from previous studies on the investigated data sets (Ohlenforst et al., 2017; Wang, Kramer, et al., 2018). In agreement with other studies investigating pupil dilation in a wide range of SNRs (Ohlenforst et al., 2017; Wendt et al., 2018; Zekveld & Kramer, 2014), in the current study, PPD was highest at –10 dB SNR corresponding to ~50% speech intelligibility, while both PPD and MPD were lowest in the +5 dB SNR condition corresponding to ~95% speech intelligibility. PPD and MPD are measures that are straightforward and easy to derive and have been used in many previous studies. However, a disadvantage of the simplicity of these measures may be that they provide less insight into the underlying processes of the pupil response (e.g., as a function of time) or are less sensitive to small differences in auditory task demand (as described earlier).

Single-Value Measures in the Frequency Domain. Unexpectedly, no significant effect of SNR, but a significant effect of luminance, was revealed on IPA, the single-value measure in the frequency domain. In contrast to the current effect of luminance, the wavelet procedure was previously promoted as a way to separate and disentangle the effect of the pupil light reflex and the task-based dilation response (Boehm-Davis & Marshall, 2003). The following factors may have contributed to the current findings. First, the effect of task difficulty on IPA was previously examined using arithmetic and sustained tasks (Boehm-Davis & Marshall, 2003; Duchowski et al., 2018). In the current study, the task was shorter and required participants to listen to sentences in noise, which might have been less cognitively challenging than performing an arithmetic task. That is, results could be influenced by both the length of the trial and the cognitive task demand. Second, Duchowski et al. (2018) reported that the resolution levels in the wavelet decomposition needs to be further validated by investigating the resolution of the wavelet decomposition required for detecting dilation reflexes in a broader range of experiments. Peysakhovich et al. (2015) reported an effect of task difficulty and no effect of luminance on the ratio of high (1.4–4 Hz) and low (0–1.4 Hz) frequency content as measured in the TEPR. The work on developing IPA is ongoing, and the technique may be optimized in future studies (Duchowski et al., 2020; Krejtz et al., 2020).

Analysis of the Pupil Time Course. The cluster-based permutation test revealed an effect of both SNR and luminance on the pupil time course. As illustrated in Figure 4, this statistical test indicated effect of SNR and luminance in relatively long clusters. Significant effects were found in the time intervals between 3.2 s and 7.7 s (–20 dB vs. +5 dB SNR, Data Set A), between 3.42 s and 8 s (–10 dB vs. +5 dB SNR, Data Set A), and between 3.27 s and 6.43 s (dark vs. light, Data Set B). These intervals include both sentence and masker presentation in Data Set A and mostly masker presentation in Data Set B. Interestingly, the cluster revealed in Data Set B was slightly delayed in comparison to Data Set A, which was likely caused by the unexpected quick dilation in Data Set A (condition +5 dB; see Figure 4). In addition to the single-value measures, applying the permutation test provided more detailed information about the effects of SNR and luminance on the temporal profiles of the TEPR. In the current study, we did not specifically test whether differences occurred during the listening, rehearsal, or respond intervals. However, this type of time-related hypotheses can be tested with the permutation test (Johansson et al., 2018).

An effect of SNR and luminance was found on multiple pupil measures obtained from both time course analysis approaches (GCA polynomial terms and RPC scores). Differences in the effect of SNR on the investigated measures (see Table 4) may suggest that those pupil measures tap into slightly different aspects of the pupil dilation response. GCA polynomial terms highlighted changes in the pupil morphology on the group level, whereas RPC scores indicated individual contributions to the three components identified in the TEPR. In alignment with Kuchinsky et al. (2013), our results on GCA show that the pupil response in acoustically challenging, but not impossible, conditions (~50% speech intelligibility) were significantly highest (intercept term), most elevated (linear term), and most delayed (cubic term). Similar to Wendt et al. (2018), we observed an effect of SNR on multiple GCA polynomial terms, which makes GCA a promising measure for studying pupil morphology. Unexpectedly, the quadratic term in the +5 dB SNR condition in Data Set A was positive (see Figure 8). It was likely caused by the quick dilation seen in Figure 4, meaning that the quadratic term was mainly based on the decreasing pupil size following the peak dilation. This example illustrates a limitation of GCA—the model is driven by assumptions, and it will likely not bring meaningful estimations for noncanonical TEPRs. This aspect of GCA may be a limiting factor in analyzing individual TEPRs or longer analysis windows, where the TEPR does not have a canonical form (van Rij et al., 2019; Winn, 2016).

Unlike other time course analysis approaches, PCA divided TEPRs into three different time windows (RPC loadings) and provided a measure of the similarity between individual and averaged pupil responses within indicated time windows (RPC scores). By comparing Figures 4 and 9, one can see an overlap between the RPC loadings and morphology of the pupil response. The RPC1 loading spanned the sentence presentation interval in which the pupil dilated; the RPC2 loading included the maximum dilation of the pupil size; and the RPC3 loading included the end of the masker presentation where the pupil size became smaller and stabilized. This pattern of loadings as well as the number of retained RPCs is in agreement with previous studies (Johansson et al., 2018; Verney et al., 2004). Thereby, PCA seems to provide an advantage over the cluster-based permutation test. While both analysis approaches indicated differences in the time course due to SNR and luminance, PCA provides more information (i.e., distinct components associated with specific time intervals) that might refer to different processes involved in performing the task at hand. Note that negative RPC scores (like in Data Set B) do not directly relate to pupil behavior, rather are a result of unrestricted PCA. Previous studies indicated that the RPC scores may provide a direct link to the response of the ANS (Wetzel et al., 2016; Widmann et al., 2018), or cognitive resources spent on stimuli processing and successfully performing the task (Johansson et al., 2018; Verney et al., 2004; Zellin et al., 2011). Based on the timing of RPC loadings with respect to the speech-in-noise test, we hypothesize that the PCA could reflect the listening (RPC1), processing the sentence (RPC2), and preparing for the response (RPC3) phases of performing a speech-in-noise test. However, further research would be required to test this hypothesis and to understand the role and meaning of the three RPCs in context of listening and speech perception. For example, one could perform a study including task manipulations such as SNR and a memory load manipulation to verify that RPC2 and RPC3 relate to different cognitive processes; alternatively, one could include both manipulation of SNR and luminance to investigate whether RPC1 relates to the activity of the nervous system or listening to the sentence.

General Discussion

An effect of SNR was found on most pupil measures, supporting pupillometry as a sensitive measure of listening effort during speech-in-noise testing. Our findings indicate that both overall changes in pupil size (PPD, MPD) and in pupil morphology (permutation test, GCA, PCA) may be important to consider in studying listening effort. The choice of specific pupil measure should be driven by the research question and its complexity. The inclusion of measures of pupil morphology is in line with a growing body of research looking into time course analysis approaches for pupillometry

(Kuchinsky et al., 2013; van Rij et al., 2019; Wagner et al., 2019; Wendt et al., 2018; Winn et al., 2015).

In the current study, we aimed to evaluate a broad range of pupil measures. Some were extracted directly from the pupil responses (PPD, MPD, and IPA), and others estimated through statistical modeling of the time course (GCA, PCA). To make a fair evaluation, we decided to compare the effects of SNR and luminance using LMM analyses. However, differences in the best-fitting models between measures limited the insights into the random effects of individuals as a subject factor was not included in all models. Furthermore, a comparison of the effects of SNR and luminance on the pupil measures as estimated by a best-fitting model required multiple statistical tests. Our findings indicated that an adequate interpretation of pupillometry data may depend on selecting the pupil measures and statistical tests most suitable to answer the specific research questions.

There were several general limitations in the current study, which could be addressed in future research. First, the data sets differed in their quality as well as in the number of responses acquired per condition. Second, the investigated pupil data were recorded with a relatively low sampling frequency and using relatively short recordings and an uncontrolled response time. Those recordings were adequate for the evaluation of MPD and PPD measures; however, their characteristics could have had an impact on the frequency and time course analyses. Third, the interaction between SNR and luminance was not systematically manipulated in the analyzed data sets, which made it difficult to compare the effects of SNR and luminance in a direct way. Fourth, the current study aimed for a fair evaluation of a range of pupil measures and not on interpreting each pupil measure separately. Thus, further research would be needed to further understand value of the specific measures, for example, RPC loadings and scores, as well as the relation between various pupil measures and the cognitive processes they may reflect.

Conclusions

Using various analysis approaches, the current study investigated the effect of SNR and luminance on several pupil measures derived from TEPRs recorded during a speech-in-noise test. The analysis approaches included three single-value measures (PPD, MPD, and IPA) as well as three time series measures (cluster-based permutation test, GCA, PCA). Our results revealed an effect of SNR on pupil measures derived from the time domain (PPD, MPD, GCA, PCA), but not the frequency domain (IPA). Findings from time course analysis approaches (permutation test, GCA, PCA) indicated that the effect of SNR and luminance can be seen in multiple aspects of the TEPR morphology (GCA) and longer time windows (permutation test, PCA). In disagreement with previous studies, the current study was not able to support claims suggesting that the IPA would disentangle the effect of the pupil light reflex from the task-evoked pupil dilation. The current findings encourage further exploration of analyses of the pupil time course as they seem to reveal aspects of pupil responses (timing or morphology) that may not be reflected in the single-value measures.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable suggestions on how to improve the article. Besides, the authors want to thank J. H. Christensen for sharing his expertise on statistical modeling and support with the LMM analyses and G. Keidser for a language check of the article. Finally, the authors would like to thank J. H. M. van Beek for his technical support in creating and validating preprocessing and analysis software used in the study.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie-Sklodowska-Curie grant agreement no. 765329 (part of the HEAR-ECO765329 project).

ORCID iDs

Patrycja Książek https://orcid.org/0000-0001-8931-453X
Lorenz Fiedler https://orcid.org/0000-0002-7892-6917

Supplemental material

Supplemental material for this article is available online.

References

Aldrich, E. (2020). *wavelets: Functions for computing wavelet filters, wavelet transforms and multiresolution analyses, R package*. https://cran.r-project.org/package=wavelets

Ayasse, N. D., Lash, A., & Wingfield, A. (2017). Effort not speed characterizes comprehension of spoken sentences by older adults with mild hearing impairment. *Frontiers in Aging Neuroscience*, *8*, 329. https://doi.org/10.3389/fnagi.2016.00329

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. http://dx.doi.org/10.18637/jss.v067.i01.

Bishop, C. M. (2006). Continuous Latent Variables. In *Pattern recognition and machine learning* (pp. 559–590). Springer.

Boehm-Davis, D., & Marshall, S. P. (2003). *Understanding and measuring cognitive workload: A coordinated multidisciplinary approach participation of women and girls in math and science View project NextGen-DataComm View project*. https://www.researchgate.net/publication/235214708

Costello, A. B. and Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation 10*. https://doi.org/10.7275/jyj1-4868

Davson, H. (1990). The pupil. In *Physiology of the eye*. Palgrave. https://doi.org/10.1007/978-1-349-09997-9_26.

Demberg, V., & Sayeed, A. (2016). The frequency of rapid pupil dilations as a measure of linguistic processing difficulty. *PLoS One*, *11*(1), e0146194. https://doi.org/10.1371/journal.pone.0146194

Dimitrijevic, A., Smith, M. L., Kadis, D. S., & Moore, D. R. (2019). Neural indices of listening effort in noisy environments. *Scientific Reports*, *9*(1), 11278. https://doi.org/10.1038/s41598-019-47643-1

Duchowski, A. T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., Raubal, M., and Giannopoulos, I. (2018). The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)* (pp. 1–13). Association for Computing Machinery, https://doi.org/10.1145/3173574.3173856

Duchowski, A. T., Krejtz, K., Gehrer, N. A., Bafna, T., and Bækgaard, P. (2020). The low/high index of pupillary activity. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)* (pp.1–12). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376394

Francis, A. L., & Oliver, J. (2018). Psychophysiological measurement of affective responses during speech perception. *Hearing Research*, *369*, 103–119. https://doi.org/10.1016/j.heares.2018.07.007

Frossard, J. and Renaud, O. (2019). *permuco: Permutation Tests for Regression, (Repeated Measures) ANOVA/ANCOVA and Comparison of Signals*. R package version 1.1.0. https://CRAN.R-project.org/package = permuco

Holman, J. A., Drummond, A., Hughes, S. E., & Naylor, G. (2019). Hearing impairment and daily-life fatigue: A qualitative study. *International Journal of Audiology*, *58*, 408–416. https://doi.org/10.1080/14992027.2019.1597284

Hothorn, T., Bretz, F. and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biome Journal, 50*, 346–363. https://doi.org/10.1002/bimj.200810425

Jackson, I., & Sirois, S. (2009). Infant cognition: Going full factorial with pupil dilation. *Developmental Science*, *12*(4), 670–679. https://doi.org/10.1111/j.1467-7687.2008.00805.x

Johansson, R., Pärnamets, P., Bjernestedt, A., & Johansson, M. (2018). Pupil dilation tracks the dynamics of mnemonic interference resolution. *Scientific Reports*, *8*(1), 4826. https://doi.org/10.1038/s41598-018-23297-3

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*, 1583–1585. https://doi.org/10.1126/science.154.3756.1583

Kayser, J., & Tenke, C. E. (2003). Optimizing PCA methodology for ERP component identification and measurement: Theoretical rationale and empirical evaluation. *Clinical Neurophysiology*, *114*(12), 2307–2325. https://doi.org/10.1016/S1388-2457(03)00241-4

Keidser, G., Seeto, M., Rudner, M., Hygge, S., & Rönnberg, J. (2015). On the relationship between functional hearing and depression. *International Journal of Audiology*, *54*, 653–664. https://doi.org/10.3109/14992027.2015.1046503

Koelewijn, T., de Kluiver, H., Shinn-Cunningham, B. G., Zekveld, A. A., & Kramer, S. E. (2015). The pupil response reveals increased listening effort when it is difficult to focus attention. *Hearing Research*, *323*, 81–90. https://doi.org/10.1016/j.heares.2015.02.004

Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing*, *33*, 291–300. https://doi.org/10.1097/AUD.0b013e3182310019

Koelewijn, T., Zekveld, A. A., Festen, J. M., Rönnberg, J., & Kramer, S. E. (2012). Processing load induced by informational masking is related to linguistic abilities. *International Journal of Otolaryngology*, *2012*, 865731. https://doi.org/10.1155/2012/865731

Kramer, S. E., Kapteyn, T. S., Festen, J. M., & Kuik, D. J. (1997). Assessing aspects of auditory handicap by means of pupil dilatation. *International Journal of Audiology*, *36*, 155–164. https://doi.org/10.3109/00206099709071969

Kramer, S. E., Kapteyn, T. S., & Houtgast, T. (2006). Occupational performance: Comparing normally-hearing and hearing-impaired employees using the Amsterdam Checklist for Hearing and Work. *International Journal of Audiology*, *45*(9), 503–512. https://doi.org/10.1080/14992020600754583

Krejtz, K., Żurawska, J., Duchowski, A. T., & Wichary, S. (2020). Pupillary and microsaccadic responses to cognitive effort and emotional arousal during complex decision making. *Journal of Eye Movement Research*, *13*(5). https://doi.org/10.16910/jemr.13.5.2

Książek, P. (2017). *Statistical modeling of pupil curves to quantify differences in processing effort on group- and individual-level*. Technical University of Denmark (DTU), Hearing Systems, Department of Electrical Engineering.

Kuchinsky, S. E., Ahlstrom, J. B., Vaden, K. I., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2013). Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology*, *50*(1), 23–34. https://doi.org/10.1111/j.1469-8986.2012.01477.x

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. http://dx.doi.org/10.18637/jss.v082.i13

Liu, Y., Rodenkirch, C., Moskowitz, N., Schriver, B., & Wang, Q. (2017). Dynamic lateralization of pupil dilation evoked by locus coeruleus activation results from sympathetic, not parasympathetic, contributions. *Cell Reports*, *20*(13), 3099–3112. https://doi.org/10.1016/j.celrep.2017.08.094

Livingston, G., Sommerlad, A., Orgeta, V., Costafreda, S. G., Huntley, J., Ames, D., … Mukadam, N. (2017). Dementia prevention, intervention, and care. *The Lancet*, *390*, 2673–2734. https://doi.org/10.1016/S0140-6736(17)31363-6

Loewenfeld, I. E. (1999). The pupil: Anatomy, physiology, and clinical applications: By Irene E. Loewenfeld. 1999. Oxford: Butterworth-Heinemann. Price pound180. Pp. 2278. ISBN 0-750-67143-2. *Brain, 124*, 1881–1883. https://doi.org/10.1093/brain/124.9.1881

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Marshall, S. P. (2002). The index of cognitive activity: Measuring cognitive workload. In *Proceedings of the IEEE 7th Conference on Human Factors and Power Plants*. https://doi.org/10.1109/HFPP.2002.1042860.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475–494. https://doi.org/10.1016/j.jml.2007.11.006

Nachtegaal, J., Smit, J. H., Smits, C., Bezemer, P. D., Van Beek, J. H. M., Festen, J. M., & Kramer, S. E. (2009). The association between hearing status and psychosocial health before the age of 70 years: Results from an internet-based national survey on hearing. *Ear and Hearing*, *30*, 302–312. https://doi.org/10.1097/AUD.0b013e31819c6e01

Ohlenforst, B., Wendt, D., Kramer, S. E., Naylor, G., Zekveld, A. A., & Lunner, T. (2018). Impact of SNR, masker type and noise reduction processing on sentence recognition performance and listening effort as indicated by the pupil dilation response. *Hearing Research*, *365*, 90–99. https://doi.org/10.1016/j.heares.2018.05.003

Ohlenforst, B., Zekveld, A. A., Lunner, T., Wendt, D., Naylor, G., Wang, Y., Versfeld, N. J., Kramer, S. E. (2017). Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation. *Hearing Research*, *351*, 68–79. https://doi.org/10.1016/j.heares.2017.05.012

Percival, D. B., & Walden, A. T. (2000). *Wavelet methods for time series analysis*. Cambridge University Press. https://doi.org/10.1017/CBO9780511841040

Peysakhovich, V., Causse, M., Scannella, S., & Dehais, F. (2015). Frequency analysis of a task-evoked pupillary response: Luminance-independent measure of mental effort. *International Journal of Psychophysiology*, *97*(1), 30–37. https://doi.org/10.1016/j.ijpsycho.2015.04.019

Peysakhovich, V., Vachon, F., & Dehais, F. (2017). The impact of luminance on tonic and phasic pupillary responses to sustained cognitive load. *International Journal of Psychophysiology*, *112*, 40–45. https://doi.org/10.1016/j.ijpsycho.2016.12.003

Pichora-Fuller, M. K. (2003). Cognitive aging and auditory information processing. *International Journal of Audiology*, *42*, 26–32. https://doi.org/10.3109/14992020309074641

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., … Wingfield, A. (2016). Hearing impairment and cognitive energy. *Ear and Hearing*, *37*, 5S–27S. https://doi.org/10.1097/AUD.0000000000000312

Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D., R Core Team. (2020). nlme: Linear and nonlinear mixed effects models. *R Package Version 3.1-148*. https://cran.r-project.org/package = nlme

R Core Team. (2020). A *language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.r-project.org

Reilly, J., Kelly, A., Kim, S. H., Jett, S., & Zuckerman, B. (2018). The human task-evoked pupillary response function is linear: Implications for baseline response scaling in pupillometry. *Behavior Research Methods*, *51*, 865–878. https://doi.org/10.3758/s13428-018-1134-4

Revelle, W. (2020). *psych: Procedures for personality and psychological research*, Northwestern University. https://CRAN.R-project.org/package = psych Version = 2.0.12.

Rstudio Team. (2016). RStudio: Integrated development for R. RStudio, Inc., Boston MA. *RStudio*. https://doi.org/10.1007/978-3-642-20966-6

Sassenhagen, J., & Draschkow, D. (2019). Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology*, *56*(6), e13335. https://doi.org/10.1111/psyp.13335

Siegle, G. J., Ichikawa, N., & Steinhauer, S. (2008). Blink before and after you think: Blinks occur prior to and following cognitive load indexed by pupillary responses. *Psychophysiology*, *45*, 679–687. https://doi.org/10.1111/j.1469-8986.2008.00681.x

Steinhauer, S. R., & Hakerem, G. (1992). The pupillary response in cognitive psychophysiology and schizophrenia. In D. Friedman & G. E. Bruder (Eds.), *Annals of the New York Academy of Sciences: Vol. 658. Psychophysiology and experimental psychopathology: A tribute to Samuel Sutton* (pp. 182–204). New York Academy of Sciences.

Steinhauer, S. R., Siegle, G. J., Condray, R., & Pless, M. (2004). *Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing*. *International Journal of Psychophysiology*, *52*, 77–86. https://doi.org/10.1016/j.ijpsycho.2003.12.005

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Lawrence Erlbaum Associates.

van Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., & Wood, S. N. (2019). Analyzing the time course of pupillometric data. *Trends in Hearing*, *23*, 1–22. https://doi.org/10.1177/2331216519832483

Verney, S. P., Granholm, E., & Marshall, S. P. (2004). Pupillary responses on the visual backward masking task reflect general cognitive ability. *International Journal of Psychophysiology*, *52*(1), 23–36. https://doi.org/10.1016/j.ijpsycho.2003.12.003

Versfeld, N. J., Daalder, L., Festen, J. M., & Houtgast, T. (2000). Method for the selection of sentence materials for efficient measurement of the speech reception threshold. *The Journal of the Acoustical Society of America*, *107*, 1671. https://doi.org/10.1121/1.428451

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin and Review*, *11*, 192–196. https://doi.org/10.3758/BF03206482

Wagner, A. E., Nagels, L., Toffanin, P., Opie, J. M., & Başkent, D. (2019). Individual variations in effort: Assessing pupillometry for the hearing impaired. *Trends in Hearing*, *23*, 1–18. https://doi.org/10.1177/2331216519845596

Wang, Y., Kramer, S. E., Wendt, D., Naylor, G., Lunner, T., & Zekveld, A. A. (2018). The pupil dilation response during speech perception in dark and light: The involvement of the parasympathetic nervous system in listening effort. *Trends in Hearing*, *22*, 1–11. https://doi.org/10.1177/2331216518816603

Wang, Y., Naylor, G., Kramer, S. E., Zekveld, A. A., Wendt, D., Ohlenforst, B., & Lunner, T. (2018). Relations between self-reported daily-life fatigue, hearing status, and pupil dilation during a speech perception in noise task. *Ear and Hearing*, *39*(3), 573–582. https://doi.org/10.1097/AUD.0000000000000512

Wang Y., Zekveld A. A., Naylor G., Ohlenforst B., Jansma E. P., Lorens A., Lunner, T. & Kramer, S. E. (2016). Parasympathetic nervous system dysfunction, as identified by pupil light reflex, and its possible connection to hearing impairment. *PLoS ONE, 11*(4), e0153566. https://doi.org/10.1371/journal.pone.0153566

Wang, Y., Zekveld, A. A., Wendt, D., Lunner, T., Naylor, G., & Kramer, S. E. (2018). Pupil light reflex evoked by light-emitting diode and computer screen: Methodology and association with need for recovery in daily life. *PLoS One*, *13*, e0197739. https://doi.org/10.1371/journal.pone.0197739

Wendt, D., Koelewijn, T., Książek, P., Kramer, S. E., & Lunner, T. (2018). Toward a more comprehensive understanding of the impact of masker type and signal-to-noise ratio on the pupillary response while performing a speech-in-noise test. *Hearing Research*, *369*, 67–78. https://doi.org/10.1016/j.heares.2018.05.006

Wetzel, N., Buttelmann, D., Schieler, A., & Widmann, A. (2016). Infant and adult pupil dilation in response to unexpected sounds. *Developmental Psychobiology*, *58*(3), 382–392. https://doi.org/10.1002/dev.21377

Widmann, A., Schröger, E., & Wetzel, N. (2018). Emotion lies in the eye of the listener: Emotional arousal to novel sounds is reflected in the sympathetic contribution to the pupil dilation response and the P3. *Biological Psychology*, *133*, 10–17. https://doi.org/10.1016/j.biopsycho.2018.01.010

Wingfield, A., Amichetti, N. M., & Lash, A. (2015). Cognitive aging and hearing acuity: Modeling spoken language comprehension. *Frontiers in Psychology, 6, 684*. https://doi.org/10.3389/fpsyg.2015.00684

Winn, M. (2016). Rapid release from listening effort resulting from semantic context, and effects of spectral degradation and cochlear implants. *Trends in Hearing*, *20*, 1–17. https://doi.org/10.1177/2331216516669723

Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2015). The impact of auditory spectral resolution on listening effort revealed by pupil dilation. *Ear and Hearing*, *36*(4), e153–e165. https://doi.org/10.1097/AUD.0000000000000145

Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky, S. E. (2018). Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started. *Trends in Hearing, 22*, 1–32, 2331216518800869. https://doi.org/10.1177/2331216518800869

Zekveld, A. A., Koelewijn, T., & Kramer, S. E. (2018). The pupil dilation response to auditory stimuli: Current state of knowledge. *Trends in Hearing*, *22*, 1–25, 2331216518777174. https://doi.org/10.1177/2331216518777174

Zekveld, A. A., & Kramer, S. E. (2014). Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology*, *51*(3), 277–284. https://doi.org/10.1111/psyp.12151

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and Hearing*, *31*(4), 480–490. https://doi.org/10.1097/AUD.0b013e3181d4f251

Zellin, M., Pannekamp, A., Toepel, U., & van der Meer, E. (2011). In the eye of the listener: Pupil dilation elucidates discourse processing. *International Journal of Psychophysiology*, *81*(3), 133–141. https://doi.org/10.1016/j.ijpsycho.2011.05.009